

# DOCUMENT RESUME

ED 186 443

TM 800 060

AUTHOR Zane, Thomas: Hursh, Daniel  
 TITLE Verification of Reliability and Validity of a Behavior Rating Scale.  
 PUB DATE [ Sep 79]  
 NOTE 21p.; Paper presented at the Annual Meeting of the American Psychological Association (87th, New York, NY, September 1-5, 1979).  
 EDRS PRICE MF01/PC01 Plus Postage.  
 DESCRIPTORS \*Behavior Rating Scales; \*Check Lists; \*Parent Child Relationship; \*Test Reliability; Test Validity

## ABSTRACT

A procedure was devised to assess the degree of reliability and validity of a behavior rating scale checklist used to evaluate parents' training skills with handicapped infants. Reliability was tested by independent observers viewing videotapes of training sessions and filling out checklist ratings of the parents' behaviors. Validity was assessed by observers viewing videotapes of training sessions, recording each training behavior exhibited by the parent as correct or incorrect, and then comparing these results to results of checklist ratings for the same training session. The degree of reliability was consistently high for one of two observer pairs, and the degree of validity was high for both pairs, in that the results obtained by the checklist corresponded with the results obtained by the detailed frequency counts. The results indicated that the checklist seemed to be easy to use as well as being an accurate assessment device. (Author/CTM)

\*\*\*\*\*  
 \* Reproductions supplied by EDRS are the best that can be made \*  
 \* from the original document. \*  
 \*\*\*\*\*

ED186443

U.S. DEPARTMENT OF HEALTH,  
EDUCATION & WELFARE  
NATIONAL INSTITUTE OF  
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

# Verification of Reliability and Validity of a Behavior

## Rating Scale

Thomas Zane  
Department of Psychology  
University of Massachusetts  
Tobin Hall  
Amherst, Ma. 01003

Dr. Daniel Hursh  
Educational Psychology Department  
West Virginia University  
Morgantown, W.V. 26506

PERMISSION TO REPRODUCE THIS  
MATERIAL HAS BEEN GRANTED BY

T. ZANE

TO THE EDUCATIONAL RESOURCES  
INFORMATION CENTER (ERIC)

TM 800060

## Verification of Reliability and Validity of a Behavior Rating Scale

A procedure was devised to assess the degree of reliability and validity of a behavior rating scale checklist used to evaluate parents' training skills. Reliability was tested by independent observers viewing videotapes of training sessions and filling out checklist ratings of the parents' behaviors. Validity was assessed by observers viewing videotapes of training sessions, recording each training behavior exhibited by the parent as correct or incorrect, and <sup>then</sup> comparing these results to results of checklist ratings for the same training session. The degree of reliability was consistently high for one of two observer pairs, and the degree of validity was high for both pairs, in that the results obtained by the checklist corresponded with the results obtained by the detailed frequency counts. The results indicated that the checklist seemed to be easy to use as well as being an accurate assessment device.

## Verification of Reliability and Validity of a Behavior Rating Scale

Many different systems for recording human behavior have been devised. These observational instruments range from simple rating scales, which typically consist of generally defined categories of behaviors, to detailed observational systems, which require exact counting of all responses at different intervals. The advantages of the rating scale format are that this type of instrument usually is not difficult for staff to learn to use, and it does not require a great deal of time to implement. On the other hand, a potentially serious disadvantage with rating scales is that due to the vagueness of the scale, results obtained may not be reliable and valid, necessary characteristics of any recording system. If a checklist scale of behaviors could meet reliability and validity standards, then it would be an observational instrument of simple design, easily used, and which one could have confidence in the results.

A checklist rating scale was created to assess the quality of training exhibited by parents of handicapped infants (copies of the checklist and instructions for use may be obtained by contacting the authors). This checklist allowed assessment of parents' training abilities in six areas: instruction, prompt, non-word cue, reinforcement, model, and ignoring. Each of these training skill categories were specifically defined. Staff used the checklist in the following manner. As a parent worked on a task with the child, the staff person observed the training. After training ended, the observer immediately filled out the checklist by rating the proportion of training behaviors, per category, exhibited correctly according to the pre-determined definitions. For example, the staff person determined that out of all the times the parent gave instructions, only half of these instances were

2

correctly exhibited as determined by the definition for the use of instructions. Thus, the staff person marked "half" on the rating scale.

Once this checklist scale was created, it was necessary to determine if it met reliability and validity standards. First, it was necessary to determine if independent observers could agree in their scoring of the training behaviors of the same parent during the same session (reliability). Also, it was necessary to determine if results from any checklist actually portrayed what really occurred during that session (validity). Reliability was tested by two observers viewing a videotape of a training session (involving parent and child) and independently filling out the checklist ratings of the parent's behavior. Reliability was defined in terms of the proportion of checklist categories which the two observers scored exactly the same. Validity was assessed by first having two observers view a videotape of a training session and record each training behavior exhibited by the parent as either correct or incorrect. This yielded an exact frequency count of the parent's training behavior. The results of this frequency count were compared to the results of checklist ratings for the same training session. The extent to which each category was scored exactly the same way on both the checklist and frequency counts determined the degree of validity.

#### Method

For reliability and validity assessment two pairs of observers were used, hereafter identified as Pairs 1 and 2. Each pair worked at different time of the day.

Videotapes of 50 different training activities involving a parent working with the child were used for assessment purposes. Confidentiality of clients was protected by (a) obtaining parent's permission to make videotapes, (b) never identifying parent or child by name to the observers, and (c) notifying the observers that their work--

3  
was to remain confidential. Each activity was approximately 3-5 minutes in length. The activities were viewed by the observers in a random order.

Reliability assessment Reliability was computed by dividing the number of categories that the two observers rated the same by the total number of categories rated, multiplied by 100. Reliability was always computed using the initial ratings for a training activity, before observers discussed the ratings.

The observers were taught to use the checklist by instructions and practice, in the following order:

1. The history, rationale, and purpose were explained;
2. The definitions of each category were reviewed and explained;
3. Each observer completed the "example lesson #2" sheet;
4. A "training tape" (videotape consisting of parents working with their children, illustrating different types of training behaviors) was viewed and the observers recorded which behaviors were exhibited;
5. The procedure for completing the actual ratings on the checklist was discussed;
6. The first taped activity was then rated by each pair of observers. First, the definitions of one of the categories of that activity were discussed. Then the activity was played one time, and when it ended the observers independently rated only that category. Then the definitions of a second category were discussed and the activity played again. When it finished the observers independently rated this second category. This procedure of evaluating one category at a time continued for all categories. After all were rated, the reliability of agreement was computed. A discussion followed concerning the ratings and any disagreements. The same activity was viewed in the manner described above until 80-100% reliability was obtained between the two observers on each category.
7. For the next two activities, observers rated two categories at one time. Thus, the definitions of two categories were discussed and then the tape of that activity was played one time. Observers independently rated the two categories. Then the definitions of the next two categories were discussed, the tape played once again, and independent ratings made. This procedure continued until all of the categories were rated. Reliability of agreement was computed.



The same activity was viewed in the manner described above until 80-100% reliability was obtained between the two observers for each category.

8. Beginning with the fourth activity, the observers were required to rate all categories from just one viewing of the activity. The definitions for each category were first discussed. Then the tape was played one time, after which the observers independently rated each category. The reliability was computed and a discussion of any discrepancies followed. If the initial reliability was less than 80-100% for any category, the activity was played again until it met that criterion.

Reliability assessment continued in this manner until the rate of percent agreement consisted of a stable level or downward (negative) trend, over five consecutive activities viewed during one 60 minute work session. At this point validity assessment began.

Validity assessment Validity was determined by the agreement between ratings of a checklist and the frequency of behaviors actually exhibited by parents during a session. To assess validity the observers first learned to compute an exact frequency count of each correct and incorrect training behavior exhibited by the parent.

When computing the frequency count, the tape of an activity was re-played as often as either observer requested. Observers used stop watches to time durations of and intervals between behaviors. The observers marked each instance of a training behavior as either correct or incorrect, according to the definitions of the categories. In this phase as well, reliability of agreement between scores obtained by these frequency counts of the two observers was computed. Reliability scores were computed using the initial counts of each observer. Any category with a percent reliability of 80-100% from the initial viewing was considered reliably measured. However, the observers were required to watch the activity again and do another frequency count of behaviors within any category that did not yield this percent

reliability agreement.

Once reliability in each category was 80-100%, each observer scored the results of the frequency count in terms of the checklist ratings. For example, if by the frequency count it was shown that 57% of all Instructions were used correctly, an observer rated the Instruction category as "half" (34-67%).

The observers were taught to perform the frequency count evaluation by instruction and practice in the following order:

1. The rationale and purpose was explained;
2. The frequency count data form was explained;
3. The frequency count of the first activity consisted of the two observers viewing the tape and discussing and recording what was observed. In other words, the observers did not record independently.
4. Once the observers agreed on the recording of the first activity, the same tape was played again. This time the observers independently counted the behaviors. After the activity was completed, the reliability of agreement between the observers was computed. A discussion followed concerning any differences in the frequency counts. The same activity was used until there was 80-100% reliability for each category.
5. Beginning with the next activity the observers independently computed a frequency count of parent behaviors. Only discussion of the definitions of the categories was allowed before the computing began. Reliability was computed for the initial scoring. If any category yielded less than 80-100% reliability, the tape was played again and frequency counts made until the percent agreement reached this criterion.

Use of the frequency counts continued in this manner. The activities used for viewing were those activities on which the other pair of observers had obtained a 100% reliability with their checklist ratings.

To determine the actual validity assessment, the ratings obtained by the frequency count done by one pair of observers were compared to the ratings obtained by the checklists done by the other pair of observers, all on the same activity. The percent agreement between the



6

two instrument ratings was computed by dividing the number of categories that had the same ratings on both instruments, divided by the total number of categories rated, multiplied by 100.

## Results

### Reliability Assessment

Pair 1 required approximately 5 hours of instruction and practice before they began evaluating activities using the checklist, rating all categories at once. The data concerning the percent of reliability of Pair 1 are found in Figure 1. The percent agreement using the checklist ranged from 0-100%; with a mean of 58.4% and a median of 30.0%. The rate of agreement stabilized in a downward (negative) trend after 17 activities.

Pair 2 began evaluating activities using the checklist and rating all categories at once after approximately 4 hours of instruction and practice. The data concerning the percent reliability agreement of this pair of observers are found in Figure 2. The percent agreement using the checklist ranged from 0-100% with a mean of 71.8% and a median of 60.0%. The rate of agreement stabilized in a downward (negative) trend after 32 activities.

-----  
Insert Figures 1 and 2 about here  
-----

### Validity Assessment

Pair 1 required approximately 2 hours to learn how to compute the frequency count and to begin counting behaviors in activities. Pair 1 used the frequency count on a total of five activities. The reliability between

these two observers ranged from 62-88%, with a mean of 61.0% (see Figure 1). Pair 2 was instructed on the use of the frequency count for approximately 3 hours. Their reliability using the frequency count on three activities ranged from 79.4%-100% with a mean of 89.7% (see Figure 2).

The reliability between the coded activities done by one pair and the checklist ratings of the same activities by the other pair ranged from 80-100%, with a mean of 91.3% (Figure 3).

-----  
Insert Figure 3 about here  
-----

#### Discussion

These data seem to indicate that the checklist was a reliable and valid measuring instrument. The percent agreement between the two observers of Pair 1 failed to consistently fall within the 80-100% range. However, the scores were close to that level, and it is quite probable with further training these observers could produce high reliability. The other team of observers consistently yielded reliable scores.

Comparing the checklist ratings with the frequency count data yielded validity within the acceptable 80-100% range. The results of the checklists appeared to accurately reflect the quality of the parent training behaviors being exhibited during the sessions (as defined by the definitions of the categories).

When using the checklist, of all the times that two observers agreed on a rating, that ratings was usually either an "all" or "none". In other words, most of the time two observers agreed in the checklist ratings, they were rating either "all" or "none". Even though this may suggest that observers were just guessing in their ratings and merely checking either the low or high extreme of the rating, this does not seem to be

the case. These same categories were rated "all" or "none" when computing the frequency count data, which confirmed the accuracy of the checklist ratings. Also, on the few categories which were reliably rated "few", "half", and "most" by two observers using checklists, the exact same ratings were received using the frequency count. Thus, it appears that the checklist ratings were indeed accurate representations of the parents' training behaviors.

# All Categories at One Time

Frequency Count

Percent Agreement

100  
90  
80  
70  
60  
50  
40  
30  
20  
10  
0

One category  
at one time

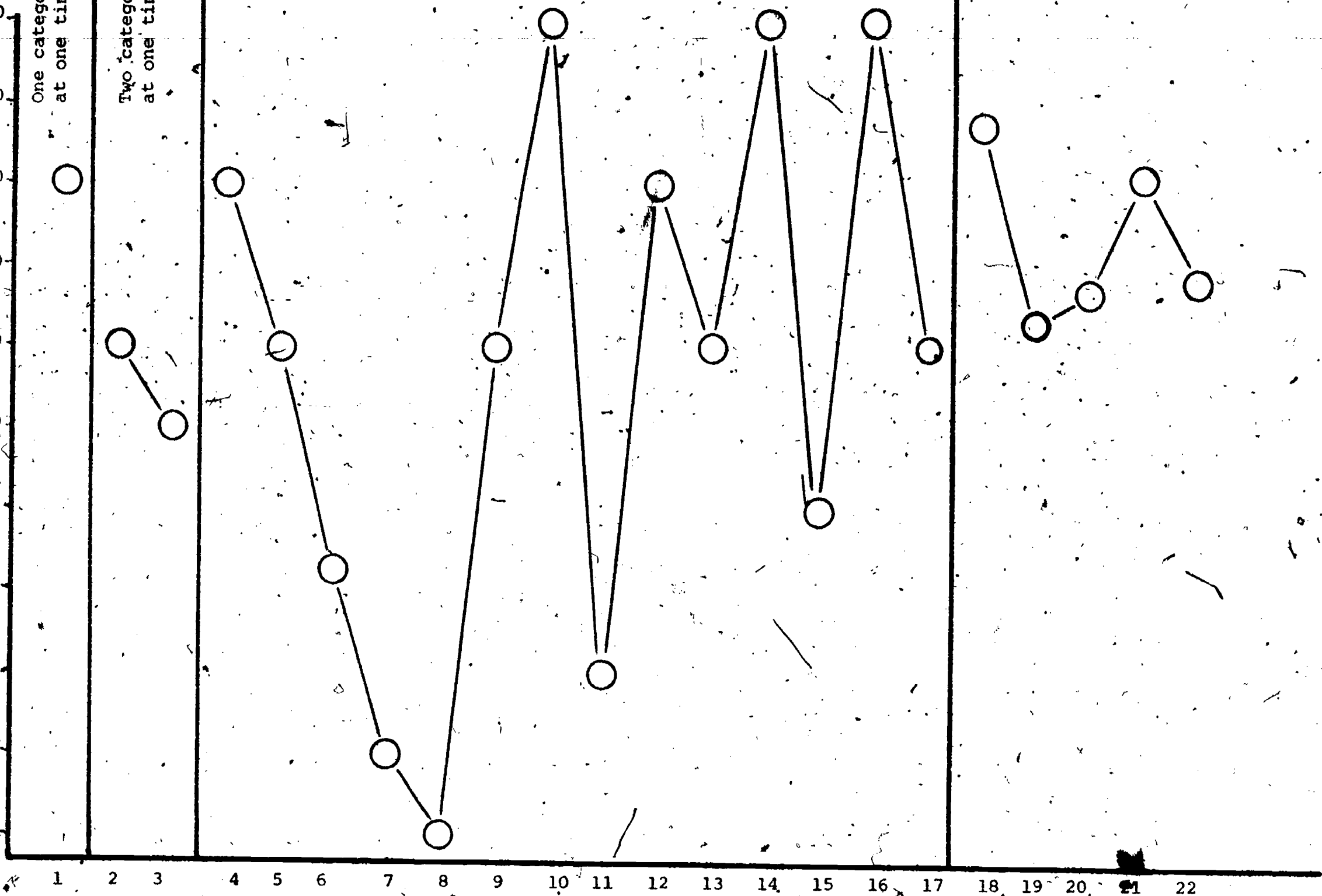
Two categories  
at one time

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Activities

Figure 1.

12



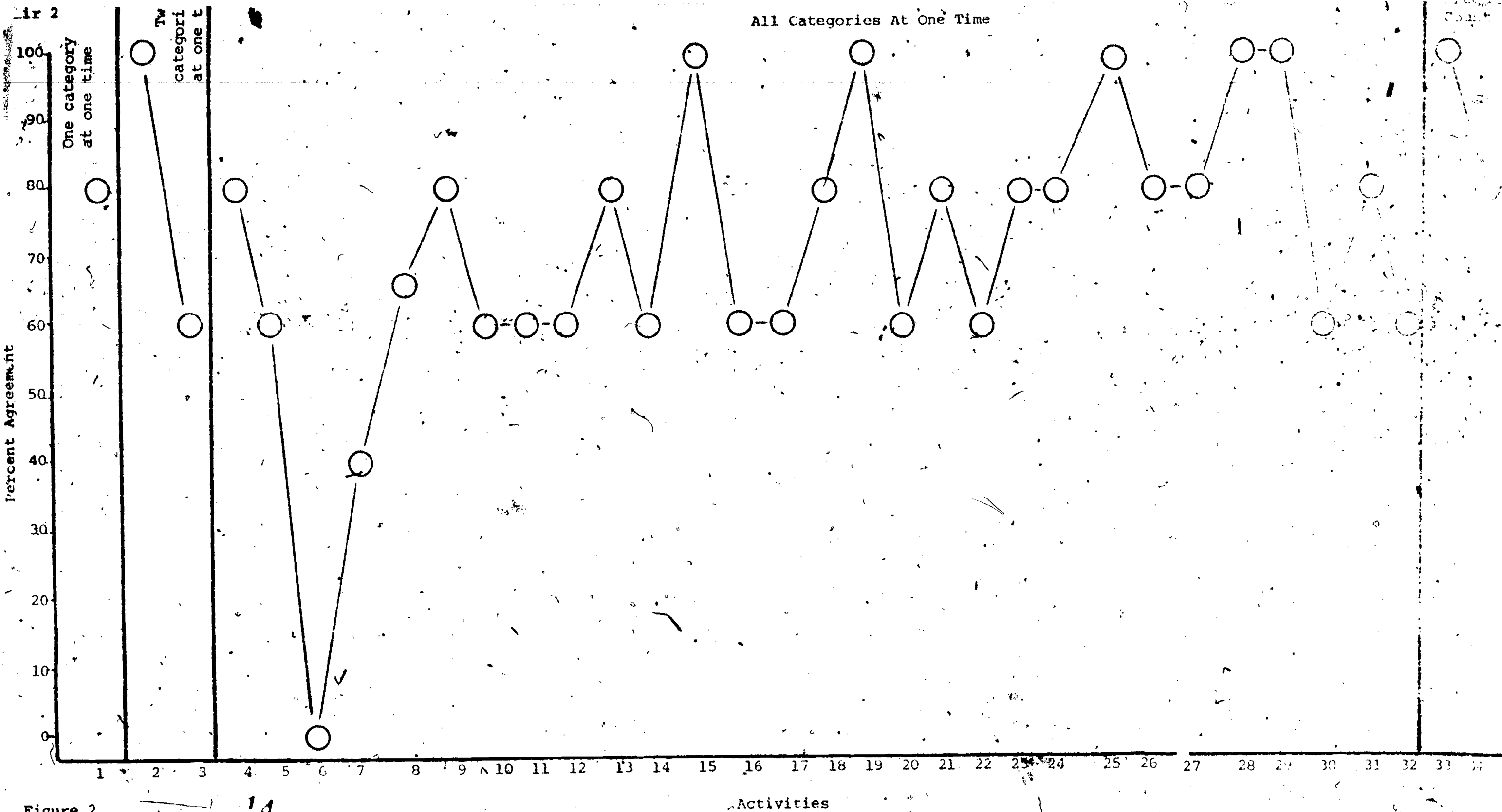
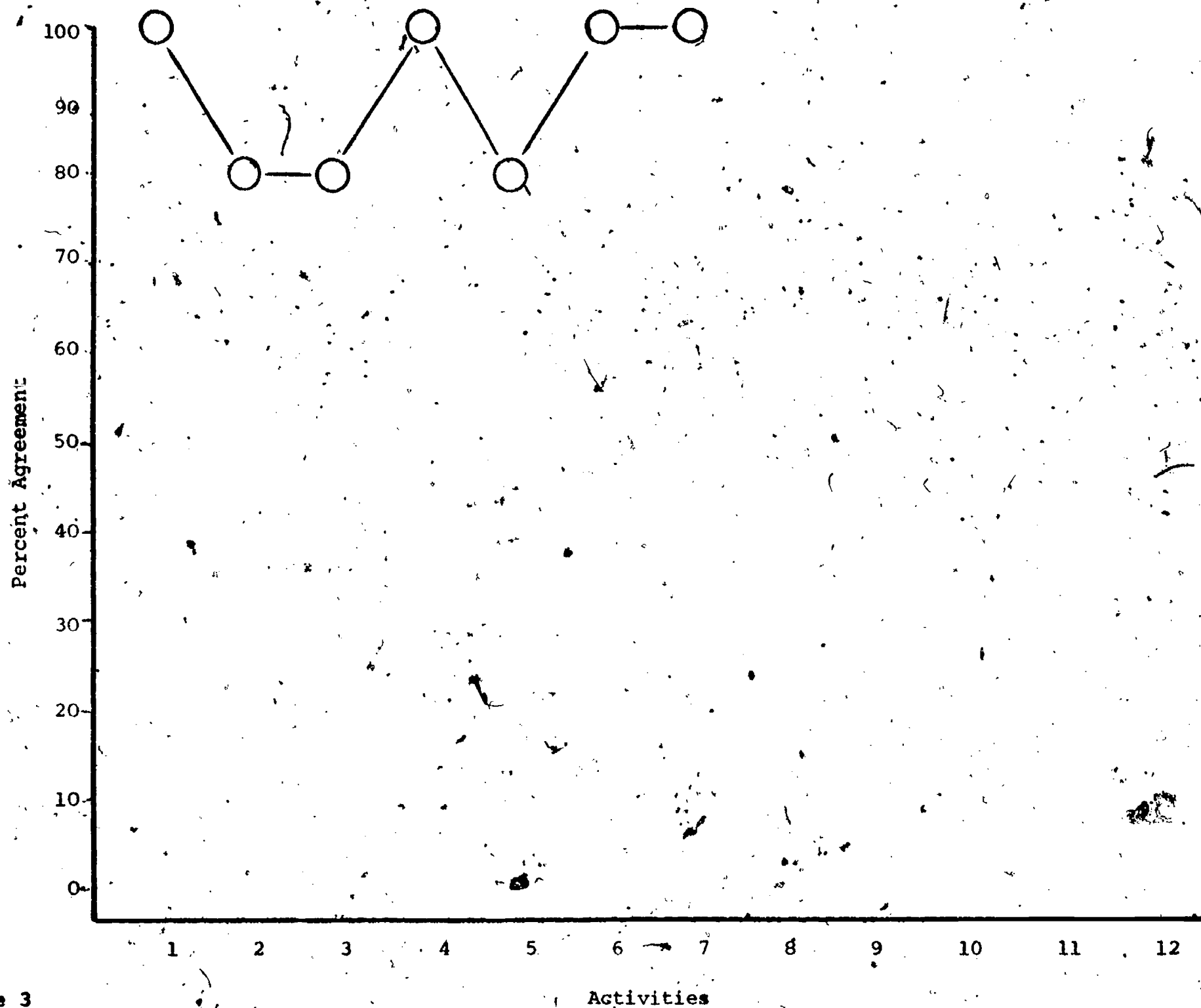


Figure 2

14

Reliability Between Frequency Counts  
and Checklist Ratings





## CHECKLIST

1. Fill in each box for each technique. There are 3 measures: (1) Frequency of technique being used; (2) Quality of the use of each technique based on attached definitions, and (3) Additional Problems observed in the use of the technique. Note: Whenever "non-given" is checked go to next technique.

A. INSTRUCTIONS: None given, none needed \_\_\_\_\_ None give, but should have \_\_\_\_\_

Frequency: Was this technique used enough?

too few \_\_\_\_\_ just enough \_\_\_\_\_ too many \_\_\_\_\_

Quality: When this technique was observed, was it used appropriate according to definitions?

None (0%) \_\_\_\_\_ Few (1-33%) \_\_\_\_\_ Half (34-67%) \_\_\_\_\_  
Most (68-99%) \_\_\_\_\_ All (100%) \_\_\_\_\_

Additional Problems:

talks too slow \_\_\_\_\_ talks too fast \_\_\_\_\_ talks too soft \_\_\_\_\_  
talks too loud \_\_\_\_\_ other \_\_\_\_\_

B. PROMPTS: None given, none needed \_\_\_\_\_ None given but should have \_\_\_\_\_

Frequency: Was this technique used enough?

too few \_\_\_\_\_ just enough \_\_\_\_\_ too many \_\_\_\_\_

Quality: When this technique was observed, was it used appropriate according to definitions?

None (0%) \_\_\_\_\_ Few (1-33%) \_\_\_\_\_ Half (34-67%) \_\_\_\_\_  
Most (68-99%) \_\_\_\_\_ All (100%) \_\_\_\_\_

Additional Problems:

Stops prompt when child resists \_\_\_\_\_ Too slow to give prompt \_\_\_\_\_  
Too fast to give prompt \_\_\_\_\_ Other \_\_\_\_\_

## Family and Infant Learning Program

Child \_\_\_\_\_

DATE \_\_\_\_\_

Parent Trainer \_\_\_\_\_

Home Trainer \_\_\_\_\_

Activity \_\_\_\_\_

The following definitions of appropriate teaching techniques should be read and completed before observing a training session. Definitions may be modified or eliminated according to individual parent-child needs (to modify fill in space provided; to eliminate, draw a line through undesired part of definition).

Instruction: (1) verbalizations that specify or cue target response (example: "pick up the ball", sweetie, over here");  
(2) no more than \_\_\_\_\_ consecutive instructions;  
(3) modifications \_\_\_\_\_

Prompt: (1) full or partial (circle 1 or 2 ) physical guidance to perform target response (example: grasping child's hand and putting it on the ball  
(2) must occur within 3 seconds of an instruction if child doesn't do  
(3) modifications \_\_\_\_\_

Non-Word Cue: (1) motioning, gesturing, non-word sounds that cue target response (example: making a "come here" gesture with hands to cue the child to crawl toward parent);  
(2) each must be no more than \_\_\_\_\_ seconds durations;  
(3) must occur within 3 seconds of an instruction if child doesn't do  
(4) modifications \_\_\_\_\_

Reinforcement: (1) positive comments, gestures, tone of voice, physical for target response (example: parent hugging child and says "you did it!");  
(2) must occur within 2 seconds of target response;  
(3) must occur for \_\_\_\_\_ seconds;  
(4) modifications \_\_\_\_\_

Models: (1) performing the target response so child can imitate (example: parent clapping hands when the target response is clapping hands)  
(2) no more than \_\_\_\_\_ consecutive models;  
(3) must occur within 3 seconds of an instruction if child doesn't do  
(4) modifications \_\_\_\_\_

Ignoring: (1) removing all physical and social contact (example: parent turning head away from child while child is behaving inappropriately);  
(2) must never occur after target or appropriate response;  
(3) must occur within 3 seconds of the "inappropriate" behavior;  
(4) modifications \_\_\_\_\_

The Second Pause Rule: Those techniques which are interrupted by a one-second pause and then repeated again are considered as one technique per pause.  
Example: for Instructions: "Come over here", one second pause, "Come here"= instructions)

C. NON-WORD CUES: None given, none needed \_\_\_\_\_ None given, but should have \_\_\_\_\_

Frequency: Was this technique used enough?

too few \_\_\_\_\_ just enough \_\_\_\_\_ too many \_\_\_\_\_

Quality: When this technique was observed, was it used appropriately?

None (0%) \_\_\_\_\_ Few (1-33%) \_\_\_\_\_ Half (34-67%) \_\_\_\_\_

Most (68-99%) \_\_\_\_\_ All (100%) \_\_\_\_\_

Additional Problems:

too distracting \_\_\_\_\_ unclear \_\_\_\_\_ wrong kind \_\_\_\_\_

other \_\_\_\_\_

D. MODELS: None given, none needed \_\_\_\_\_ None given, but should have \_\_\_\_\_

Frequency: Was this technique used enough?

too few \_\_\_\_\_ just enough \_\_\_\_\_ too many \_\_\_\_\_

Quality: When this technique was observed, was it used appropriately?

None (0%) \_\_\_\_\_ Few (1-33%) \_\_\_\_\_ Half (34-67%) \_\_\_\_\_

Most (68-99%) \_\_\_\_\_ All (100%) \_\_\_\_\_

Additional Problems:

too slow \_\_\_\_\_ too fast \_\_\_\_\_ child not watching \_\_\_\_\_

too long \_\_\_\_\_ Other \_\_\_\_\_

E. REINFORCEMENT: None given, no correct child behavior occurred \_\_\_\_\_

None given, but should have \_\_\_\_\_

Frequency: Was this technique used enough?

too few \_\_\_\_\_ just enough \_\_\_\_\_ too many \_\_\_\_\_

Quality: When this technique was observed, was it used appropriately?

None (0%) \_\_\_\_\_ Few (1-33%) \_\_\_\_\_ Half (34-67%) \_\_\_\_\_

Most (68-99%) \_\_\_\_\_ All (100%) \_\_\_\_\_

II. Summary of teaching:

A. Were SMALL STEPS rewarded appropriately?

None given, no correct child behavior occurred \_\_\_\_\_

Yes \_\_\_\_\_ No \_\_\_\_\_

B. Use of teaching techniques:

	OK	Needs Work	
		Frequency	Quality
1. Instructions	_____	_____	_____
2. Prompts	_____	_____	_____
3. Non-Word Cues	_____	_____	_____
4. Models	_____	_____	_____
5. Reinforcement	_____	_____	_____
6. Ignoring	_____	_____	_____

Additional Problems:

Short Duration \_\_\_\_\_ Long Duration \_\_\_\_\_ Unexcited Delivery \_\_\_\_\_

None or little social praise \_\_\_\_\_ Given after inappropriate

behavior \_\_\_\_\_ Other \_\_\_\_\_

F. IGNORING: None given, no inappropriate child behavior occurred \_\_\_\_\_

None given, but should have \_\_\_\_\_

Frequency: Was this technique used enough?

Too few \_\_\_\_\_ Just enough \_\_\_\_\_ Too many \_\_\_\_\_

Quality: When this technique was observed, was it used appropriately?

None (0%) \_\_\_\_\_ Few (1-33%) \_\_\_\_\_ Half (34-67%) \_\_\_\_\_

Most (68-99%) \_\_\_\_\_ All (100%) \_\_\_\_\_

Additional Problems /

Too long \_\_\_\_\_ Too short \_\_\_\_\_ Continued Attention to Child \_\_\_\_\_

Other \_\_\_\_\_